浙江大学

硕 士 研 究 生 读 书 报 告

题目 4D Gaussian Splatting for Real-Time Dynamic Scene Rendering 读书报告

作者姓名 胡杭锦

作者学号 22351079

指导教师 李启雷

学科专业 电子信息（软件工程）

所在学院 软件学院

提交日期 2024 年 1 月

Reading Report of 4D Gaussian Splatting for Real-Time Dynamic Scene Rendering

A Dissertation Submitted to

Zhejiang University

in partial fulfillment of the requirements for

the degree of

Master of Engineering

Major Subject: Software Engineering

Advisor: Qilei Li

By

Hangjin Hu

Zhejiang University, P.R. China

2024

摘要

新视图合成(NVS)是 3D 视觉领域的一项关键任务，在 VR、AR 和电影制作等领域得到了广泛应用。其中，一项重要但具有挑战性的任务就是实现动态场景的表示和渲染。在这篇论文中，作者提出使用 4D 高斯分布（4D-GS）作为动态场景的整体表示,而不是对每个单独的帧应用 3D-GS 实现了实时动态场景渲染，保证了高训练和存储效率。在 4D-GS 中，作者还提出了一种包含 3D 高斯和 4D 神经体素的新型显式表示。受 HexPlane 启发，提出了一种分解的神经体素编码算法，可以从 4D 神经体素有效地构建高斯特征，然后应用轻量级 MLP 来预测新时间戳下的高斯变形。作者提出的 4D-GS 方法在 RTX 3090 GPU 上实现了高分辨率、800×800 分辨率下 82 FPS 的实时渲染，实现 SOTA。

关键词：新视图合成，4D-GS，动态场景，实时渲染

Abstract

New View Synthesis (NVS) is a key task in the field of 3D vision and has been widely used in fields such as VR, AR and film production. Among them, an important but challenging task is to realize the representation and rendering of dynamic scenes. In this paper, the author proposes to use 4D Gaussian distribution (4D-GS) as the overall representation of dynamic scenes, instead of applying 3D-GS to each individual frame to achieve real-time dynamic scene rendering, ensuring high training and storage efficiency. In 4D-GS, the authors also propose a new explicit representation containing 3D Gaussians and 4D neural voxels. Inspired by HexPlane, we propose a decomposed neural voxel encoding algorithm that can effectively construct Gaussian features from 4D neural voxels, and then apply lightweight MLP to predict Gaussian deformation under new timestamps. The 4D-GS method proposed by the author achieves real-time rendering of 82 FPS at high resolution and 800×800 resolution on RTX 3090 GPU, achieving SOTA.

Keywords：NVS, 4D-GS, Dynamic Scenes, Real-Time Rendering

1 引言

新视图合成(NVS)是 3D 视觉领域的一项关键任务，在许多应用中发挥着至关重要的作用，例如：VR、AR 和电影制作。NVS 旨在从场景的任何所需视点或时间戳渲染图像，通常需要根据多个 2D 图像准确地对场景进行建模。动态场景在真实场景中非常常见，渲染很重要但具有挑战性，因为复杂的运动需要使用空间和时间稀疏输入进行建模。

NeRF[1]通过用隐式函数表示场景在合成新颖的视图图像方面取得了巨大成功。引入体积渲染技术来连接 2D 图像和 3D 场景。然而，原始的 NeRF 方法承受着巨大的训练和渲染成本。尽管一些 NeRF 变体[2-8]将训练时间从几天缩短到几分钟，但渲染过程仍然存在不可忽略的延迟。

最近的 3D 高斯分布(3D-GS)[9]通过将场景表示为 3D 高斯，显着地将渲染速度提高到实时水平。原始 NeRF 中繁琐的体积渲染被高效的可微分泼溅技术取代，它直接将 3D 高斯点投影到 2D 平面上。3D-GS 不仅具有实时渲染速度，而且能够更明确地表示场景，从而更容易操纵场景表示。

然而，3D-GS 专注于静态场景。将其作为 4D 表示扩展到动态场景是一个合理、重要但困难的课题。关键的挑战在于根据稀疏输入对复杂的点运动进行建模。3DGS 通过用点状高斯表示场景来保留自然几何先验。一种直接有效的扩展方法是在每个时间戳构建 3D 高斯[10]，但存储/内存成本会成倍增加，特别是对于长输入序列。本文目标是构建紧凑的表示，同时保持训练和渲染效率，即 4D 高斯分布（4DGS）。为此，作者建议通过有效的高斯变形场网络来表示高斯运动和形状变化，该网络包含时空结构编码器和极小的多头高斯变形解码器，仅维护一组规范 3D 高斯。对于每个时间戳，规范 3D 高斯将通过高斯变形场变换到具有新形状的新位置，变换过程代表高斯运动和变形。与单独建模每个高斯的运动不同，时空结构编码器可以连接不同的相邻 3D 高斯以预测更准确的运动和形状变形，然后将变形后的 3D 高斯分布用于渲染根据时间戳的图像。

2 方法

2.1 方法概述

4D-GS 的整体技术路线如下所示。

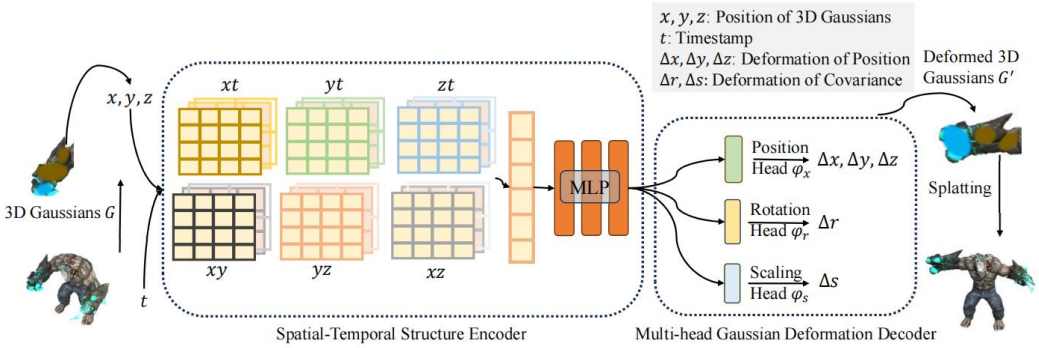


Figure 3. The overall pipeline of our model. Given a group of 3D Gaussians $\mathcal{G}$, we extract the center coordinate of each 3D Gaussian $\mathcal{X}$ and timestamp t to compute the voxel feature by querying multi-resolution voxel planes. Then a tiny multi-head Gaussian deformation decoder is used to decode the feature and get the deformed 3D Gaussians $\mathcal{G}'$ at timestamp t. The deformed Gaussians are then splatted to the rendered image.

图 1 4D-GS 技术路线

从整体上来看， 3D-GS 为了实现高质量动态场景的实时渲染，主要包括以下三个贡献。

1. 通过对高斯运动和高斯形状随时间的变化进行建模，提出了一种具有高效高斯变形场的高效 4D 高斯喷射框架。

2. 提出了一种多分辨率编码方法来连接附近的 3D 高斯并通过高效的时空结构编码器构建丰富的 3D 高斯特征。

3. 4D-GS 实现动态场景实时渲染，合成数据集在 800×800 分辨率下可达 82 FPS，在 1352×1014 分辨率下真实数据集可达 30 FPS，同时保持与之前状态相当或更好的性能最先进的 (SOTA) 方法，并显示出在 4D 场景中编辑和跟踪的潜力。

2.2 4D Gaussian Splatting 框架

如图 1 所示，给定视图矩阵 M = [R, T]，时间戳 t，4D 高斯分布框架包括了 3D 高斯 G 和高斯变形场网络 F。然后通过微分溅射渲染新视角图像 $\hat{I}$，遵循 $\hat{I} = S(M, G')$，其中 G’= Δ G+G。

具体来说，3D 高斯的变形 Δ G 是由高斯变形场网络 Δ G = F(G, t)，其中时空结构编码器 H 可以对 3D 高斯函数的时间和空间特征进行编码 $f_d = H(G,t)$，多头高斯变形解码器 D 可以对特征进行解码并预测每个 3D 高斯函数的变形 Δ G =

D(f)，则可以引入变形的 3D 高斯 G’。4D 高斯分布的渲染过程如图 2 (c) 所示。4D 高斯分布将原始 3D 高斯 G 转换为给定时间戳 t 的另一组 3D 高斯 G'，保持了中提到的差分分布的有效性。

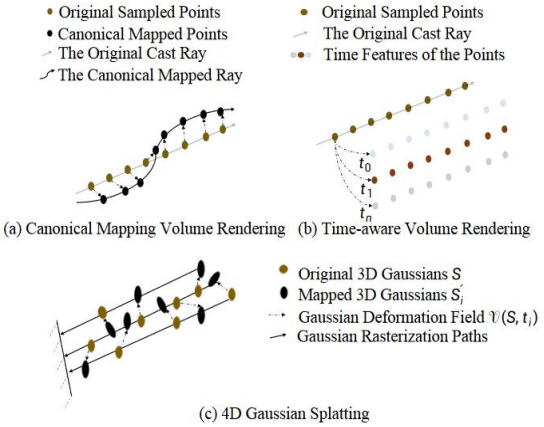


图 2 不同动态场景渲染方法图示

2.3 高斯变形场

高斯变形场包括高效的时空结构编码器 H 和高斯变形解码器 D，用于预测每个 3D 高斯的变形。

2.3.1 时空结构编码器

相邻 3D 高斯总是共享相似的空间和时间信息。为了有效地建模 3D 高斯特征，作者引入了一种高效的时空结构编码器 H，包括多分辨率 HexPlane R(i, j)和微小的 MLP φd。虽然普通 4D 神经体素非常消耗内存，但作者采用 4D K-Planes 模块将 4D 神经体素分解为 6 个平面。某个区域中的所有 3D 高斯分布都可以包含在边界平面体素中，并且高斯变形也可以编码在附近的时间体素中。具体来说，时空结构编码器 H 包含 6 个多分辨率平面模块 Rl(i, j)和一个微小的 MLP φd，即 H(G, t) = {Rl(i, j), φd|(i, j)) ∈ {(x, y), (x, z), (y, z), (x, t), (y, t), (z, t)}, l ∈ {1, 2}}。位置 X = (x, y, z) 是 3D 高斯 G 的平均值。每个体素模块由 R(i, j) ∈ Rh×lNi×lNj 定义，其中 h 代表特征的隐藏暗度，N 表示体素网格的基本分辨率，l 等于上采样比例。这需要在考虑时间信息的同时对 6 个 2D 体素平面内的 3D 高斯信息进行编码。计算单独体素特征的公式如下：

$$f_h = \bigcup_l \prod \text{interp}(R_l(i,j)),$$
$$(i,j) \in \{(x,y),(x,z),(y,z),(x,t),(y,t),(z,t)\}.$$

fh ∈ Rh*l 是神经体素的特征。插值表示用于查询位于网格 4 个顶点的体素

特征的双线性插值。然后通过一个微小的 MLP φd 通过 fd = φd(fh)合并所有特征。

2.3.2 多头高斯变形解码器

当 3D 高斯的所有特征都被编码后，作者可以使用多头高斯变形解码器 D = {φx, φr, φs} 计算任何所需的变量。采用单独的 MLP 来计算位置变形ΔX = φx(fd)、旋转Δr = φr(fd)和缩放Δs = φs(fd)。那么，变形特征（X′，r′，s′）可以表示为：（X′，r′，s′）=（X+ΔX，r+Δr，s+Δs）。最后得到变形的 3D 高斯 G′={X′，s′，r′，σ，C}。

2.4 优化过程

2.4.1 3D 高斯初始化

3D 高斯可以通过运动结构 (SfM)点初始化进行良好训练。同样，4D 高斯也应该在适当的 3D 高斯初始化中进行微调。作者在初始 3000 次迭代中优化 3D 高斯函数以进行预热，然后使用 3D 高斯函数$\hat{I}$ = S(M, G)而不是 4D 高斯函数$\hat{I}$ = S(M, G')渲染图像。优化过程的图示如图 3 所示。

2.4.2 损失函数

与其他重建方法[3,9,11]类似，作者使用 L1 颜色损失来监督训练过程。还应用了基于网格的总变分损失[12,3,5,8]L_{tv}。

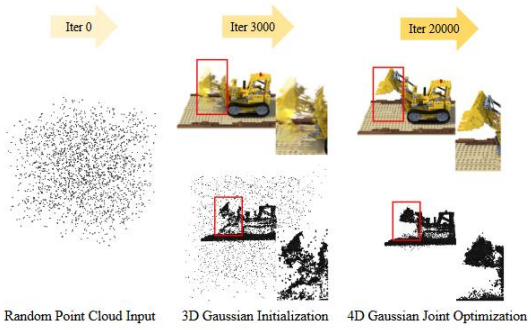


图 3 优化过程图示

3 实验与结果

3.1 数据集

作者主要基于 PyTorch 框架，并在单个 RTX 3090 GPU 中进行测试，并通过 3D-GS [9]中概述的配置微调了相关优化参数。

3.1.1 合成数据集

作者主要使用 DNeRF [11]引入的合成数据集来评估模型的性能。这些数据集是针对单眼设置而设计的，但值得注意的是每个时间戳的相机姿势接近随机生成。这些数据集中的每个场景都包含动态帧，数量从 50 到 200 不等。

3.1.2 真实世界数据集

作者使用 HyperNeRF[13]和 Neu3D[14]提供的数据集作为基准数据集来评估模型在现实场景中的性能。Nerfies 数据集是使用一台或两台相机在简单的相机运动后捕获的，而 Neu3D 的数据集是使用 15 到 20 个静态相机捕获的，涉及较长的周期和复杂的相机运动。作者使用 SfM 从 Neu3D 数据集中每个视频的第一帧计算出的点以及 HyperNeRF 中随机选择的 200 帧。

3.2 实验结果

作者主要使用各种指标来评估实验结果，包括峰值信噪比（PSNR）、感知质量测量 LPIPS [15]、结构相似性指数（SSIM）[16]及其扩展，包括结构相异性指数测量（DSSIM）、多尺度结构相似性指数（MS-SSIM）、FPS、训练时间和存储。

表 1 合成数据集评估结果

Model	PSNR(dB)↑	SSIM↑	LPIPS↓	Time↓	FPS ↑	Storage (MB)↓
TiNeuVox-B [6]	32.67	0.97	0.04	28 mins	1.5	48
KPlanes [8]	31.61	0.97	-	52 mins	0.97	418
HexPlane-Slim [4]	31.04	0.97	0.04	11m 30s	2.5	38
3D-GS [14]	23.19	0.93	0.08	10 mins	170	10
FFDNeRF [12]	32.68	0.97	0.04	-	< 1	440
MSTH [37]	31.34	0.98	0.02	6 mins	-	-
Ours	34.05	0.98	0.02	20 mins	82	18

表 2 HyperNeRF 数据集评估结果

Model	PSNR(dB)↑	MS-SSIM↑	Times↓	FPS↑	Storage(MB)↓
Nerfies [24]	22.2	0.803	∼ hours	< 1	-
HyperNeRF [25]	22.4	0.814	32 hours	< 1	-
TiNeuVox-B [6]	24.3	0.836	30 mins	1	48
3D-GS [14]	19.7	0.680	40 mins	55	52
FFDNeRF [12]	24.2	0.842	-	0.05	440
Ours	25.2	0.845	1 hour	34	61

表 3 Neu3D 数据集评估结果

Model	PSNR(dB)↑	D-SSIM↓	LPIPS↓	Time ↓	FPS↑	Storage (MB)↓
NeRFPlayer [35]	30.69	0.034	0.111	6 hours	0.045	-
HyperReel [2]	31.10	0.036	0.096	9 hours	2.0	360
HexPlane-all* [4]	31.70	0.014	0.075	12 hours	0.2	250
KPlanes [8]	31.63	-	-	1.8 hours	0.3	309
Im4D [18]	32.58	-	0.208	28 mins	∼5	93
MSTH [37]	32.37	0.015	0.056	20 mins	2(15‡)	135
Ours	31.15	0.016	0.049	40 mins	30	90

为了评估新视图合成的质量，作者针对该领域的几种最先进的方法进行了基准测试，结果总结在表 1 中。从表中数据可以发现，虽然当前的动态混合表示可以产生高质量的结果，但它们通常存在渲染速度的缺点。缺乏对动态运动部分的建模使得无法重建动态场景。相比之下，本文方法在合成数据集中享有最高的渲染质量和极快的渲染速度，同时保持极低的存储消耗和收敛时间。

此外，从真实世界数据集获得的结果显示在表 2、3 中。从表中数据可以发现，一些 NeR 方法收敛速度慢，而其他基于网格的 NeRF 方法在尝试捕捉复杂的物体细节时遇到困难。而本文方法研究了可比较的渲染质量、快速收敛，并且在室内情况下的自由视图渲染速度方面表现出色。尽管 Im4D 解决了与 4DGS 相比的高质量问题，但对多摄像头设置的需求使得单目场景建模变得困难，并且其他方法也限制了自由视图渲染速度和存储。

3.3 消融实验

表 4 消融实验

Model	PSNR(dB)↑	SSIM↑	LPIPS↓	Time↓	FPS↑	Storage (MB)↓
Ours w/o HexPlane $R_l(i,j)$	27.05	0.95	0.05	10 mins	140	12
Ours w/o initialization	31.91	0.97	0.03	19 mins	79	18
Ours w/o ϕ_x	26.67	0.95	0.07	20 mins	82	17
Ours w/o ϕ_r	33.08	0.98	0.03	20 mins	83	17
Ours w/o ϕ_s	33.02	0.98	0.03	20 mins	82	17
Ours	34.05	0.98	0.02	20 mins	82	18

- 时空结构编码器

显式 HexPlane 编码器 Rl(i, j)具有保留 3D 高斯空间和时间信息的能力，与纯显式方法相比，可以减少存储消耗。放弃这个模块，作者发现仅使用浅层 MLP ϕd 无法对各种设置下的复杂变形进行建模。图 3 表明，虽然模型的内存成本极低，但它确是以牺牲渲染质量为代价。

• 高斯变形解码器

作者提出的高斯变形解码器 D 对来自时空结构编码器 H 的特征进行解码。3D 高斯的所有变化都可以通过单独的 MLP {ϕx, ϕr, ϕs} 来解释。如表 4 所示，如果不建模 3D 高斯运动，4D 高斯就无法很好地拟合动态场景。同时，人体关节的运动在宏观上通常表现为表面细节的拉伸和扭曲。如果想要准确地模拟这些运动，3D 高斯的大小和形状也应该相应调整。否则，过度拉伸时可能会出现细节拟合不足的情况，或者无法在微观层面正确模拟物体的运动。图 4 还表明 3D 高斯的形状变形对于恢复细节很重要。

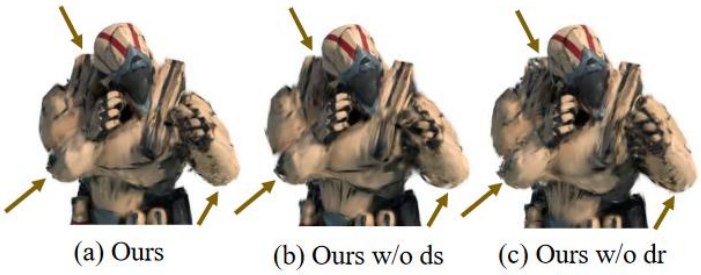


图 4 消融实验

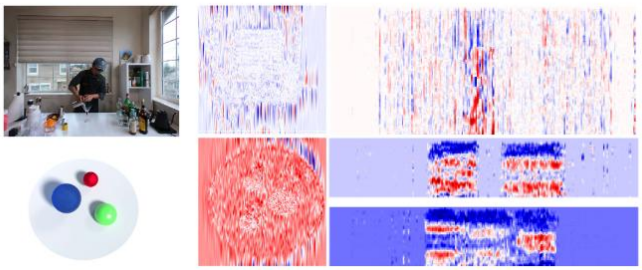
• 3D 高斯初始化

在某些没有 SfM 点初始化的情况下，直接训练 4D-GS 可能会导致收敛困难。优化 3D 高斯函数以进行热身：

(a) 让一些 3D 高斯函数留在动态部分，这可以释放 4D 高斯函数大变形学习的压力，如图 3 所示。

(b) 学习适当的 3D 高斯函数 G 并提出建议变形场更加关注动态部分，例如图 5(c)。

(c) 避免优化高斯变形网络 F 时的数值误差并保持训练过程稳定。图 3 还表明，如果在没有预热粗略阶段的情况下训练模型，渲染质量将会受到影响。



(a) The Rendered Image (b) Spatial Grid (c) Temporal Grid

图 5 多分辨率体素网格可视化

4 总结与展望

本文提出 4D 高斯喷射来实现实时动态场景渲染。构建有效的变形场网络来准确地模拟高斯运动和形状变形，其中相邻的高斯通过时空结构编码器连接。高斯之间的连接导致更完整的变形几何形状，有效避免撕脱。本文的 4D 高斯不仅可以对动态场景进行建模，还具有进行 4D 目标跟踪和编辑的潜力。

虽然 4D-GS 确实可以在许多场景中实现快速收敛并产生实时渲染结果，但仍有一些关键挑战需要解决。

1. 对数据集要求高

大的运动、背景点的缺失和不精确的相机姿势会导致优化 4D 高斯函数的困难。且对数据的预处理较为繁多，对数据集的泛化能力较差。

2. 无法分割静态和动态

在没有任何额外监督的情况下，4D-GS 还无法在单目设置下分割静态和动态高斯部分的联合运动，这仍然是一个挑战。

3. 只能处理较小场景

由于大量 3D 高斯对高斯变形场的大量查询，需要设计一种更紧凑的算法来处理城市规模的重建。

5 参考文献

[1] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. Communications of the ACM, 65(1):99–106, 2021. 1, 2

[2] Ang Cao and Justin Johnson. Hexplane: A fast representation for dynamic scenes. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 130–141, 2023. 1, 2, 4, 5, 6, 7, 12, 13

[3] Jiemin Fang, Taoran Yi, Xinggang Wang, Lingxi Xie, Xiaopeng Zhang, Wenyu Liu, Matthias Nießner, and Qi Tian. Fast dynamic radiance fields with time-aware neural voxels. In SIGGRAPH Asia 2022 Conference Papers, pages 1–9, 2022. 1, 2, 4, 5, 6, 7, 11, 12, 13

[4] Sara Fridovich-Keil, Alex Yu, Matthew Tancik, Qinhong Chen, Benjamin Recht, and Angjoo Kanazawa. Plenoxels: Radiance fields without neural networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 5501–5510, 2022.

[5] Sara Fridovich-Keil, Giacomo Meanti, Frederik Rahbæk Warburg, Benjamin Recht, and Angjoo Kanazawa. K-planes: Explicit radiance fields in space, time, and appearance. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 12479–12488, 2023. 1, 2, 4, 5, 6, 7, 12, 13

[6] Thomas Muller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. ACM Transactions on Graphics (ToG), 41(4):1–15, 2022. 1

[7] Ruizhi Shao, Zerong Zheng, Hanzhang Tu, Boning Liu, Hongwen Zhang, and Yebin Liu. Tensor4d: Efficient neural 4d decomposition for high-fidelity dynamic reconstruction and rendering. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 16632– 16642, 2023. 1, 2, 4

[8] Cheng Sun, Min Sun, and Hwann-Tzong Chen. Direct voxel grid optimization: Super-fast convergence for radiance fields reconstruction. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 5459– 5469, 2022. 1, 5

[9] Bernhard Kerbl, Georgios Kopanas, Thomas Leimk Muhler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. ACM Transactions on Graphics (ToG), 42(4):1–14, 2023. 2, 3, 4, 5, 6, 11, 12, 13

[10] Jonathon Luiten, Georgios Kopanas, Bastian Leibe, and Deva Ramanan. Dynamic 3d gaussians: Tracking by persistent dynamic view synthesis. In 3DV, 2024. 2, 3, 7, 8, 13, 14

[11] Albert Pumarola, Enric Corona, Gerard Pons-Moll, and Francesc Moreno-Noguer. D-nerf: Neural radiance fields for dynamic scenes. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 10318–10327, 2021. 2, 4, 5, 6, 11, 13

[12] Benjamin Attal, Jia-Bin Huang, Christian Richardt, Michael Zollhoefer, Johannes Kopf, Matthew O'Toole, and Changil Kim. Hyperreel: High-fidelity 6-dof video with rayconditioned sampling. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 16610–16620, 2023. 5, 6, 7

[13] Keunhong Park, Utkarsh Sinha, Peter Hedman, Jonathan T Barron, Sofien Bouaziz, Dan B Goldman, Ricardo MartinBrualla, and Steven M Seitz. Hypernerf: A higherdimensional representation for topologically varying neural radiance fields. arXiv preprint arXiv:2106.13228, 2021. 2, 4, 5, 6, 11, 12, 13

[14] Tianye Li, Mira Slavcheva, Michael Zollhoefer, Simon Green, Christoph Lassner, Changil Kim, Tanner Schmidt, Steven Lovegrove, Michael Goesele, Richard Newcombe, et al. Neural 3d video synthesis from multi-view video. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 5521–5531, 2022. 2, 4, 6, 7, 11, 12, 13

[15] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 586–595, 2018. 6

[16] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing, 13(4):600–612, 2004. 6